Journal of Modern Applied Statistical Methods

Volume 10 | Issue 1 Article 16

5-1-2011

Factors Influencing the Mixture Index of Model Fit in Contingency Tables Showing Independence

Xuemei Pan
IBM Global Business Services, xpan@us.ibm.com

C. Mitchell Dayton
University of Maryland, cdayton@umd.edu

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Pan, Xuemei and Dayton, C. Mitchell (2011) "Factors Influencing the Mixture Index of Model Fit in Contingency Tables Showing Independence," *Journal of Modern Applied Statistical Methods*: Vol. 10 : Iss. 1 , Article 16.
DOI: 10.22237/jmasm/1304223300

Journal of Modern Applied Statistical Methods
May 2011, Vol. 10, No. 1, 176-191
Copyright © 2011 JMASM, Inc.
1538 – 9472/11/$95.00

Factors Influencing the Mixture Index of Model Fit in Contingency Tables Showing Independence

Xuemei Pan
IBM Global Business Services,
Springfield, VA USA

C. Mitchell Dayton
University of Maryland,
College Park, MD USA

Several competing computational techniques for dealing with sampling zeros were evaluated when estimating the two-point mixture model index, π^*, in contingency tables under an independence assumption. Also, the performance of the estimate and associated standard errors were studied under various combinations of conditions.

Key words: Mixture index, contingency tables, sampling zeros, standard error, Monte Carlo simulation.

Introduction

Traditional methods for evaluating models for contingency table data based on Chi-square statistics or quantities derived from such statistics are not attractive in many applied research settings. According to Rudas (1998), "First, when the model is not true, a comparison of the data to what could only be expected if it were is of very little meaning; second, the actual distribution of the statistic may be very different from the reference distribution if some of the underlying assumptions are violated" (page 15). In addition, conventional methods are sensitive to sample size; often a model is rejected when fitted to a large data set even though the model may represent a reasonable summary of the data for practical purposes.

In contrast to Chi-squared tests of fit methods, which rely heavily on size of the table, sample size and actual true probabilities (Rudas, 1998), the RCL mixture index of fit proposed by Rudas, Clogg and Lindsay (1994), provides a novel way of representing goodness-of-fit for contingency tables. In contrast to classical significance tests, this index has an intuitive rationale and it does not assume a simple model that describes the entire population; the RCL index is also not sensitive to sample size like Chi-square-related quantities. Specifically, two components (subgroups) are assumed in the population. One of size $1-\pi$, where some specified model H holds true, describes the fraction of the population that is consistent with model H (e.g., independence); the other component of size π, is completely unrestricted and represents the part of the population that is outside of model H. RCL also introduced an expectation-maximization (EM) algorithm to obtain maximum likelihood estimates of π^* and derived a way to construct a lower-bound confidence-interval estimate of $\hat{\pi}^*$. As summarized by Dayton (2003), $\hat{\pi}^*$ possesses the following properties:

1. $\hat{\pi}^*$ is always located on the 0, 1 interval;
2. $\hat{\pi}^*$ is unique;

Xuemei Pan is currently a statistician at IBM Global Business Services. She earned her Ph.D. in Measurement and Statistics from University of Maryland, College Park. Her research interests include latent class modeling and model comparison procedures. Email her at: xpan@us.ibm.com. Chan Dayton is a Professor Emeritus and past Chair in the Department of Measurement & Statistics of University of Maryland, College Park. For more than 20 years, he has pursued a research interest in latent class analysis which is a specialized field within the realm of discrete mixture models. Email him at: cdayton@umd.edu.

3. $\hat{\pi}^*$ is invariant when frequencies in a contingency table are increased or decreased by a multiplicative constant.

Properties and applications of the mixture index of fit are further explored in Clogg, Rudas and Xi (1995), Xi (1996), and Clogg, Rudas and Matthews (1997). The two-point mixture index, π^*, can be applied when models are fitted to virtually any contingency table. For example, it has been applied in differential item functioning (Rudas & Zwick, 1997), latent class analysis (Dayton, 1999), regression models with normal and uniform error structures (Rudas, 1999) and logistic regression analysis (Verdes & Rudas, 2002).

Issues concerning π^* require further examination exist because they have not been studied in RCL or in other related research. In particular:

1. $\hat{\pi}^*$ is positively biased in finite samples; that is, even if H holds so that, in theory, $\pi^* = 0$, $\hat{\pi}^*$ will have expectation greater than zero for finite samples.
2. Sampling 0's can greatly affect estimation so it is useful to study the effect of using flattening constants or redefining the model by regarding sampling zeros as structural zeros.
3. Although the estimated lower confidence bound of $\hat{\pi}^*$ introduced by RCL gives inferential information that is independent of bias, it tends to be problematic when π^* is close to zero or sample size is small; thus a parametric simulation seems to be necessary to examine this measure of precision for $\hat{\pi}^*$. (As an aside, SAS code written for this study makes these analyses more accessible to researchers in various disciplines.)

Mixture Index of Fit

Suppose H represents a hypothesized probabilistic model for a frequency table and P is the true distribution for the cell proportions in the table. The two-point mixture model is defined as:

$$P = (1-\pi)\Phi + \pi\psi \qquad (1)$$

where Φ is the probability distribution implied by H, and ψ is an arbitrary, unspecified probability distribution. The mixture parameter, π, defined on the 0, 1 interval, represents the proportion of the population that cannot be described by H. Note that π is not unique and that the representation of P in equation (1) is correct for any model for any frequency table. The index of fit, π^*, however, is defined as the smallest value of π for which equation (1) holds; that is:

$$\pi^* = \inf\{\pi \mid P = (1-\pi)\phi + \pi\psi, \phi \in H\}$$

(Rudas and Zwick, 1997). Consequently, as shown by RCL, π^* is unique and represents the minimum proportion of cases that must be excluded from the frequency table in order for P to be fitted exactly by the model.

EM Algorithm and Interval Estimation

The procedure to estimate $\hat{\pi}^*$ is as follows:

1. Set the initial estimate, $\hat{\pi}^*$ to zero;
2. Obtain maximum likelihood estimates of the parameters in the components of the two-point mixture using an expectation-maximization (EM) algorithm as above, and,
3. Successively increase $\hat{\pi}^*$ by some small increment with re-estimation of the parameters at each step (e.g., .01 is been used the example below).

The value of the likelihood ratio Chi-square fit statistic, G^2, converges to zero (e.g., less than a convergence criterion set to $<10^{-5}$) and the step at which this first occurs provides the final estimate of the fit index, $\hat{\pi}^*$. (Dayton, 2003; RCL). In addition, RCL implemented this approach in their FORTRAN program, Mixit, as described in detail by Xi (1994). As shown by RCL, an appropriate lower confidence 95% bound, $\hat{\pi}_L$, is given by the value of $\hat{\pi}$ that is associated with a G^2 fit statistic equal to 2.71

(i.e., the 90th percentage point of the one-degree-of-freedom Chi-square distribution).

Sampling Zeros

According to RCL, the effect of sampling zeros on $\hat{\pi}^*$ will depend on the structure of the data as well as the suitability of model H for the data. In general, $\hat{\pi}^*$ will tend to be overestimated by a fraction that is directly related to the smaller of the observed row marginal proportion and the observed column marginal proportion pertinent to the cell with a sampling zero. Rudas and Zwick (1997) replaced zero frequencies with small positive flattening values in data from a study by Zwick, Thayer and Wingersky (1994) to investigate the sampling zero effect on the performance of π^*. Although they concluded that increasing the flattening value resulted in reducing overestimation for estimates of π^*, the effects were very small.

Structural zeros, also called logical zeros (Knoke and Burke, 1980), arise when it is logically impossible to observe positive cell counts for particular combinations of row and column variables. To demonstrate structural zeros, a typical example of the logical impossibility of observing male obstetrical patients was presented by Fienberg (1980). In practice, researchers could evaluate the variation in π^* by setting cells with no frequency to structural zeros.

Methodology

Research Design

The following aspects of the simulation were implemented:

1. Sizes of two-way contingency tables were selected: 2×2, 2×3, 2×4, 2×6, 3×3, 4×4 and 6×6. These table sizes were chosen because they provided a reasonable range of contingency table sizes in real data settings and are typical of what is found in practice.
2. Marginal distribution: evenly distributed, slightly dispersed and extremely dispersed distribution for each different table size. (Row and column total proportions for the various sized tables are shown in Figure 1.) These marginal distributions were chosen because they represented a reasonable range of different values, and the extreme marginal values were used to ensure zero cell frequencies in the observed tables.
3. Sample size for simulated contingency table: 5, 10, 20 and 30 per cell were chosen because they entailed a practical variety of sample sizes and were large enough to demonstrate a sample size effect on the mixture index of fit.
4. Techniques for zeros cells: (A) treating as sampling zeros, (B) replacing with small flattening constants (0.1, 0.5 and 1 were used to represent extremely small, moderately small and small flattening constants range), and (C) redefining model H by regarding the sampling zeros as structural zeros.
5. In each of the above scenarios, a 95% lower confidence limit based on empirically simulated $\hat{\pi}^*$ s was calculated and compared with the lower limit estimate presented by RCL.

For each table size, sample size and marginal distribution, 1,000 frequency tables were randomly generated based on the specified cumulative distribution. For example, for a 2×2 table with sample size of 10 per cell and marginal distribution {P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 9, P_{+2}=. 1}, the theoretical cumulative distribution is {0.81, 0.90, 0.99, 1}.

To generate each of the 1,000 simulated data tables, SAS code (SAS Institute, 2005) was used to generate 40 uniform random numbers on the 0, 1 and to locate them into appropriate cumulative categories (e. g., numbers less than or equal to 0.81 were placed in cell 1, 0.81; numbers between 0.81 and 0.90 in cell 2; numbers between 0.90 and 0.99 in cell 3 and the remainder in cell 4.) The value of $\hat{\pi}^*$ and associated 95% lower bound $\hat{\pi}_L$ following RCL was obtained for each generated data table; thus for each scenario, 1,000 $\hat{\pi}^*$ values and 1,000, 95% lower bound $\hat{\pi}_L$ values were generated using RCL methods. This was repeated for each of the 96 scenarios. Also for each scenario, four techniques for sampling zeros cells were

Figure 1: Row and Column Total Proportions for the Various Sized Tables

2×2 Table

{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 5, P_{+2}=. 5},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 9, P_{+2}=. 1},
{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 9, P_{+2}=. 1}.

2×3 Table

{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 8, P_{+2}=. 1, P_{+3}=. 1},
{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 33, P_{+2}=. 33, P_{+3}=. 33},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 8, P_{+2}=. 1, P_{+3}=. 1},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 33, P_{+2}=. 33, P_{+3}=. 33}.

2×4 Table

{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 25, P_{+2}=. 25, P_{+3}=. 25, P_{+4}=. 25},
{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 4 P_{+2}=. 4, P_{+3}=. 1, P_{+4}=. 1},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 25, P_{+2}=. 25, P_{+3}=. 25, P_{+4}=. 25},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 4, P_{+2}=. 4, P_{+3}=. 1, P_{+4}=. 1}.

2×6 Table

{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 167, P_{+2}=. 167, P_{+3}=. 167, P_{+4}=. 167, P_{+5}=. 167, P_{+6}=. 167},
{P_{1+}=. 5, P_{2+}=. 5, P_{+1}=. 3, P_{+2}=. 3, P_{+3}=. 1, P_{+4}=. 1, P_{+5}=. 1, P_{+6}=. 1},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 167, P_{+2}=. 167, P_{+3}=. 167, P_{+4}=. 167, P_{+5}=. 167, P_{+6}=. 167},
{P_{1+}=. 9, P_{2+}=. 1, P_{+1}=. 3, P_{+2}=. 3, P_{+3}=. 1, P_{+4}=. 1, P_{+5}=. 1, P_{+6}=. 1}.

3×3 Table

{P_{1+}=. 4, P_{2+}=. 4, P_{3+}=. 2, P_{+1}=. 4, P_{+2}=. 4, P_{+3}=. 2},
{P_{1+}=. 33, P_{2+}=. 33, P_{3+}=. 33, P_{+1}=. 33, P_{+2}=. 33, P_{+3}=. 33},
{P_{1+}=. 33, P_{2+}=. 33, P_{3+}=. 33, P_{+1}=. 4, P_{+2}=. 4, P_{+3}=. 2}.

4×4 Table

{P_{1+}=. 25, P_{2+}=. 25, P_{3+}=. 25, P_{4+}=. 25, P_{+1}=. 25, P_{+2}=. 25, P_{+3}=. 25, P_{+4}=. 25},
{P_{1+}=. 4, P_{2+}=. 4, P_{3+}=. 1, P_{4+}=. 1, P_{+1}=. 4, P_{+2}=. 4, P_{+3}=. 1, P_{+4}=. 1},
{P_{1+}=. 25, P_{2+}=. 25, P_{3+}=. 25, P_{4+}=. 25, P_{+1}=. 4, P_{+2}=. 4, P_{+3}=. 1, P_{+4}=. 1}.

6×6 Table

{P_{1+}=. 167, P_{2+}=. 167, P_{3+}=. 167, P_{4+}=. 167, P_{5+}=. 167, P_{6+}=. 167, P_{+1}=. 167, P_{+2}=. 167, P_{+3}=. 167, P_{+4}=. 167, P_{+5}=. 167, P_{+6}=. 167},
{P_{1+}=. 3, P_{2+}=. 3, P_{3+}=. 1, P_{4+}=. 1, P_{5+}=. 1, P_{6+}=. 1, P_{+1}=. 3, P_{+2}=. 3, P_{+3}=. 1, P_{+4}=. 1, P_{+5}=. 1, P_{+6}=. 1},
{P_{1+}=. 167, P_{2+}=. 167, P_{3+}=. 167, P_{4+}=. 167, P_{5+}=. 167, P_{6+}=. 167, P_{+1}=. 3, P_{+2}=. 3, P_{+3}=. 1, P_{+4}=. 1, P_{+5}=. 1, P_{+6}=. 1}.

compared: treating zero cells as sampling zeros, replacing with different small flattening constant (i.e., 0.1, 0.5 and 1), and redefining model H by regarding a sampling zero as a structural zero.

The mean of the 1,000 $\hat{\pi}^*$ values for each scenario was calculated and served as the final parameter estimate; the mean of the 1,000 $\hat{\pi}_L$ values was also computed to be the estimate 95% $\hat{\pi}_L$ using the RCL method. Because the empirical distribution of $\hat{\pi}^*$ is notably skewed

for the generated sets of 1,000 $\hat{\pi}^*$ values, the regular normal assumption cannot be used to compute the standard error and confidence interval for $\hat{\pi}^*$. Instead, 50th $\hat{\pi}^*$ value among the 1,000 values (i.e., 5th percentage point) was adopted and treated as true 95% lower bound based on empirical simulations.

Typically, $\hat{\pi}^*$ will tend to be overestimated by a fraction that is directly related to the smaller of the observed row marginal proportion and the observed column marginal proportion related to the cell with a sampling zero (RCL). As noted above, in practice, researchers could test the π^* variation by setting some to-be-ignored cells to structural zeros and resolve. This study focused this issue on any frequency tables with only one structural zero and the procedure using EM based methodology to obtain $\hat{\pi}^*$. The two-point mixture using an expectation-maximization (EM) algorithm proposed by RCL could still be applied to structural zero conditions with minor modification as follows:

1. Obtain $\hat{\pi}^*$ treating zero cell as sampling zero utilized the same procedure in EM Algorithm and Interval Estimation; in this step the entire row or column with which smaller of observed row marginal proportion and the observed column marginal proportion would result in zero in the first component, Φ, which is defined as the probability distribution designated by H.
2. Pull the proportion back from the second component, ψ, an unspecified probability distribution outside of model H for the entire row or column with zeros in component Φ at step 1.
3. Temporally cross out the other column or row that contains the zero cell but has not been forced zero at step 1.
4. Apply the same EM based procedure in the remaining contingency table while fixing all cell proportions in component 1, Φ and component 2, ψ except the row or column has frequency pulled back in step 2.
5. After iterations converge to a preset criterion,, subtract original $\hat{\pi}^*$ at step 1 with the sum of the proportion pulled back in Φ from step 4 and the final value is the estimate of $\hat{\pi}^*$ using structural zero technique.

For the other sampling zero techniques, procedures are same as sampling zeros, simply replace the zero cell with different small flattening constant (0.1, 0.5 and 1) and recall associated $\hat{\pi}^*$ based on the EM based procedures in EM Algorithm and Interval Estimation.

Simulation Details

The simulation code was written in SAS/IML version 9.1 (SAS Institute, 2005). The EM algorithm was used to calculate the mixture index of fit. Each simulation consisted of 1,000 replications with convergence criterion set to 10^{-5}. Data were randomly generated according to cumulative proportion resulting from the different combination scenarios.

The method proceeded in the following manner:

1. A sample contingency table was randomly generated based on cumulative proportion resulting from different factor combinations. (table size, sample size and marginal distribution).
2. An EM algorithm based method for mixture index of fit (RCL) was implemented. $\hat{\pi}^*$ and 95% lower bound $\hat{\pi}_L$ were generated and saved in a matrix.
3. Replicate steps 1 and 2 1,000 times, therefore 1,000 $\hat{\pi}^*$ and $\hat{\pi}_L$ were obtained and exported into an external file. Additionally, if any of the 1,000 generated contingency tables contained zero cell(s), they were replaced with different small flattening constants 0.1, 0.5 and 1, respectively, when evaluating the performance of $\hat{\pi}^*$ using flattening constants techniques.

The only difference between the structural zero and other sampling zero technique procedure is in the above-mentioned step 1. If the frequency tables generated by UNIFORM contained 1 or less than 1 frequency zero, it would proceed to step 2 otherwise it would regenerate the table until it met the requirement.

Results

Parameter Estimates and Bias

For the conditions studied, $\hat{\pi}^*$ was significantly ($p < 0.05$) positively, biased from its expected value of zero by an amount ranging from 0.02298 (2×2 table, slightly dispersed row and column marginals with sample size equals to 30 per cell) to 0.4086 (6×6 table, evenly dispersed row and column marginals with sample size equal to 5 per cell). As shown in Figures 2 and 3, for 2×2, 2×3, 2×4, 2×6 tables, as table size increases, $\hat{\pi}^*$ consistently increased (with only two exceptions) for constant sample size (5, 10, 20 and 30 per cell) and marginal distribution (evenly, slightly and extremely dispersed).

The same conclusion applies to symmetric tables: 2×2, 3×3, 4×4, 6×6. In particular, for sample sizes 5, 10, 20 and 30 per cell in evenly dispersed tables, $\hat{\pi}^*$ increased on average from 0.1252 to 0.4086; 0.096 to 0.3031; 0.0775 to 0.2242 and 0.0668 to 0.1867 for 2×2 to 6×6 tables, respectively. For sample sizes 5, 10, 20 and 30 per cell in extremely dispersed tables, $\hat{\pi}^*$ increases on average from 0.0598 to 0.03629; 0.0568 to 0.2593; 0.0476 to 0.1942 and 0.0396 to 0.1626 for a 2×2 table to a 6×6 table, respectively.

Moreover, with few exceptions, for each frequency table, as sample size increases, the bias in $\hat{\pi}^*$ significantly decreased ($p < 0.05$,). For each size contingency table, $\hat{\pi}^*$ is, on average, smallest for extremely dispersed row and column marginal distributions, and largest on average for evenly distributed row and column tables. The only exception is the 2×2 table where a slightly dispersed table contains slightly smaller $\hat{\pi}^*$ values on average than an extremely dispersed frequency table; this is in part due to a convergence problem (using less than 0.001 instead of otherwise 0.00001).

For all two-way tables, replacing zero with larger flattening values results in smaller average values of $\hat{\pi}^*$. For all extremely dispersed and most slightly dispersed (4 out of 6 scenarios) row and column marginal distributions with small sample size (5 per cell) and small table size (2×2, 2×3, 2×4, 3×3) tables, the value of $\hat{\pi}^*$ was smaller using structural zeros compared to using sampling zeros or any other replacement with positive flattening constants. Note that the techniques of replacing zero cell with flattening constants includes virtually any number of simulated zero cells for each table while the structural zero technique used in this study can only accommodate one zero cell per frequency table. Because the number of zero counts and patterns are somewhat different among these techniques, especially when encountering small sample sizes such as 5 per cell and 10 per cell, it might influence the comparison results between structural zero and using sampling zero or any other replacing with small positive flattening constants techniques.

Lower Bound Comparisons of RCL and True Estimates

The 95% lower bound estimate for $\hat{\pi}^*$ using the RCL method is generally close to the so-called true estimate based on empirical simulations. When, under some circumstances, the RCL method underestimates the lower bound value, the magnitude of underestimation is relatively small and the difference from the true estimate decreases as the sample size increases.

Similar to parameter estimators for $\hat{\pi}^*$, the true (empirical) 95% lower bound estimates of $\hat{\pi}^*$ consistently increased as table size increased within constant sample size per cell and constant marginal distribution (Figures 6 and 7). There are exceptions for 2×3 and 2×4 extremely dispersed tables with sample size 5 for which estimates remain nearly unchanged over conditions. Also in general for each frequency table, as sample size increases the 95% lower bound decreases.

Figure 2: $\hat{\pi}^*$ for Evenly Distributed Marginals

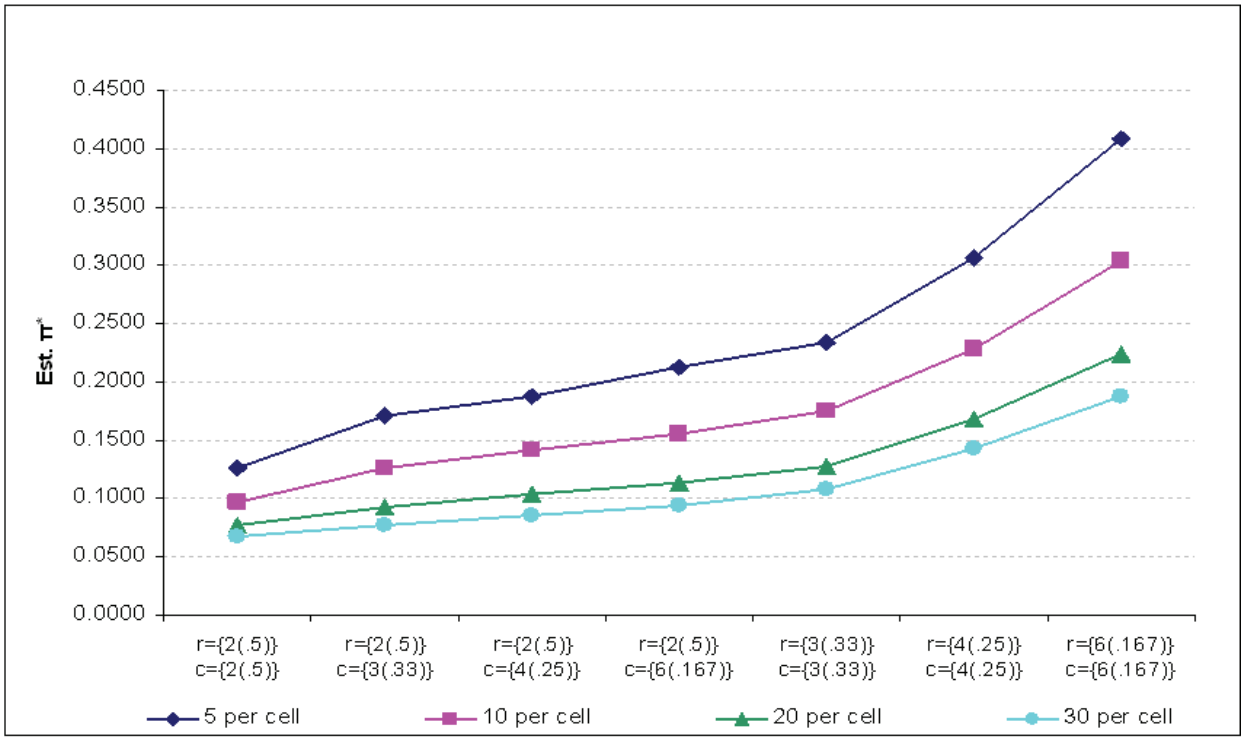


Figure 3: $\hat{\pi}^*$ for Extremely Distributed Marginals

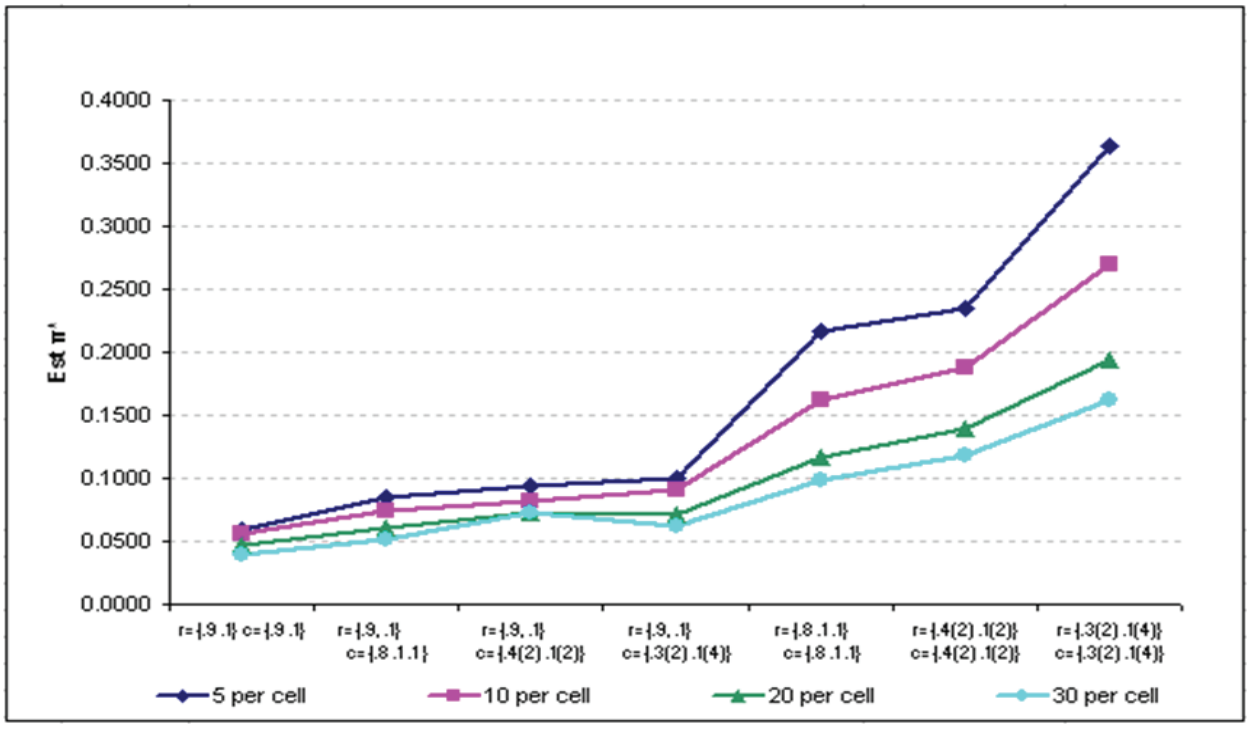


Figure 4: $\hat{\pi}^*$ Comparison in $\{P_{1+}=.9, P_{2+}=.1, P_{+1}=.8, P_{+2}=.1, P_{+3}=.1\}$

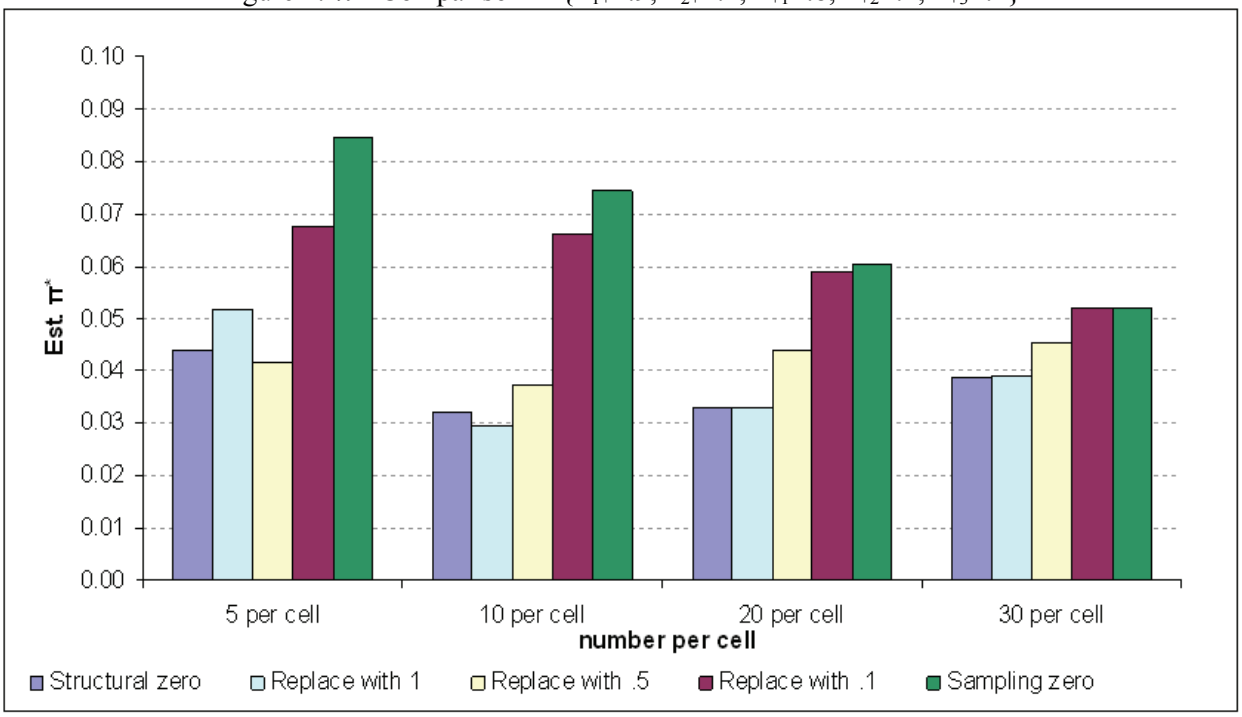


Figure 5: $\hat{\pi}^*$ Comparison in $\{P_{1+}=.9, P_{2+}=.1, P_{+1}=.4, P_{+2}=.4, P_{+3}=.1, P_{+4}=.1\}$

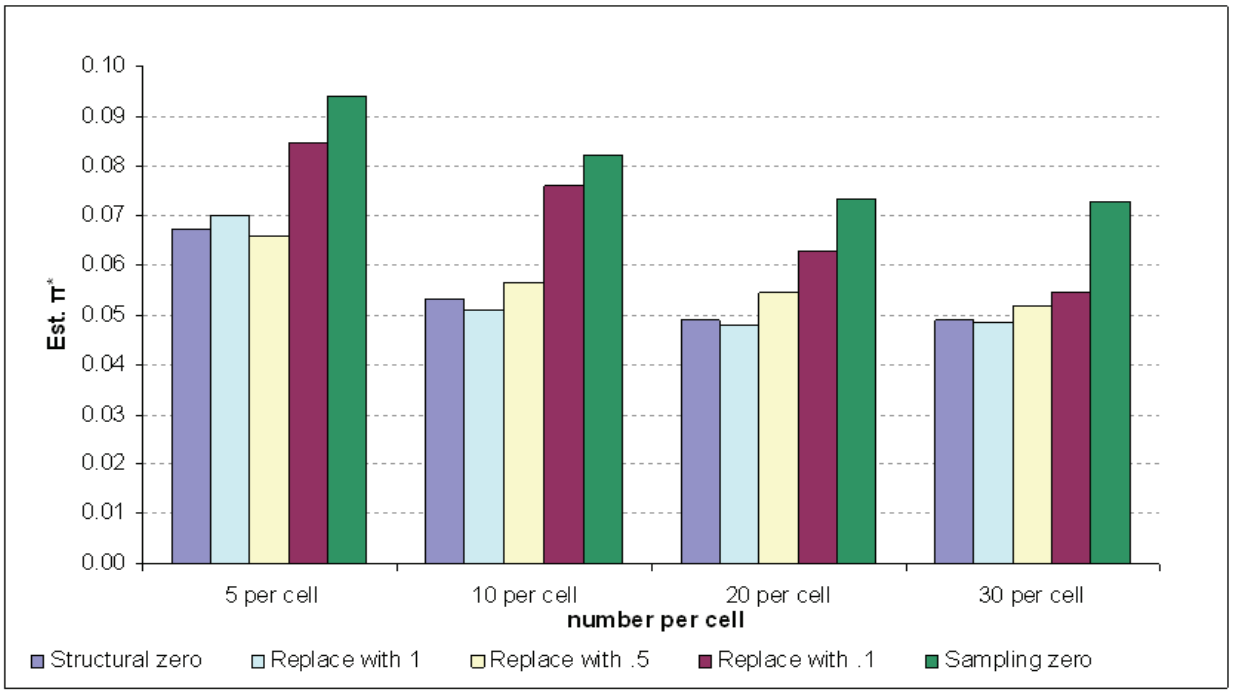


Figure 6: Empirical Simulation Based $\hat{\pi}_L$ with Evenly Distributed Marginals

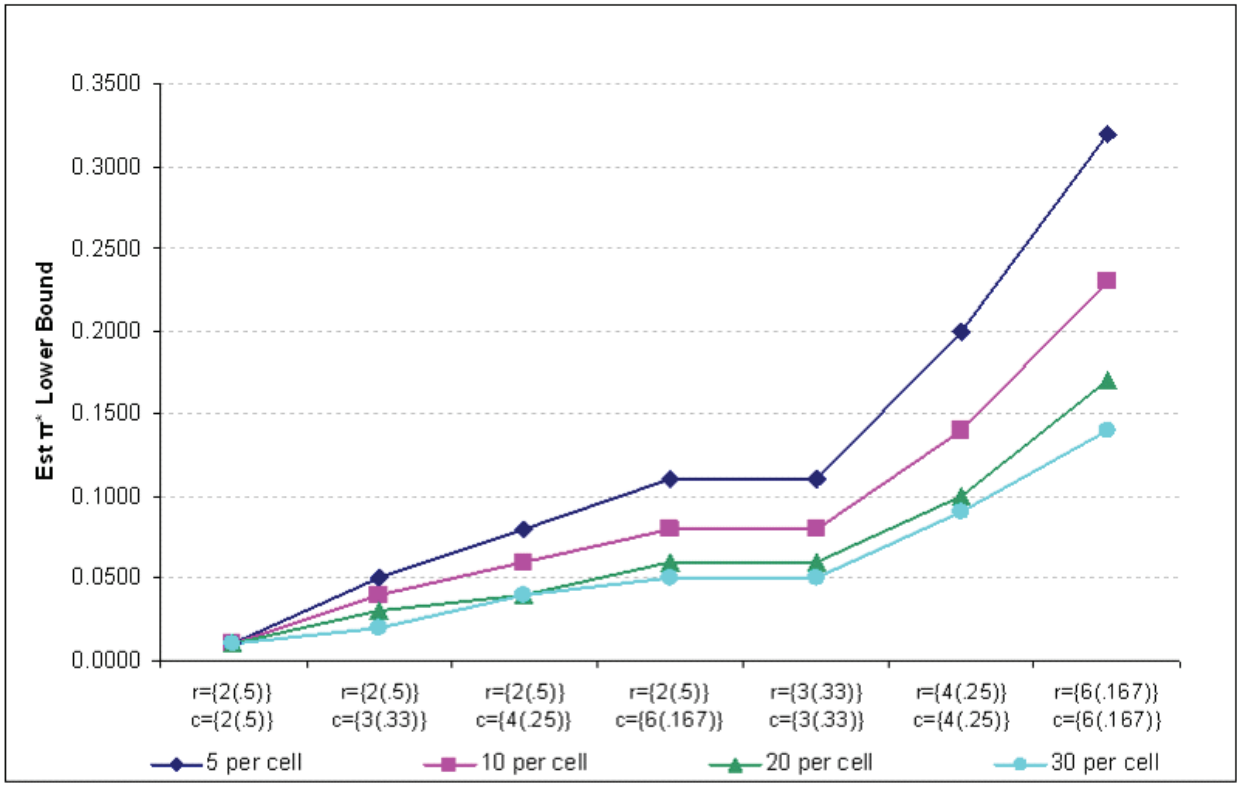
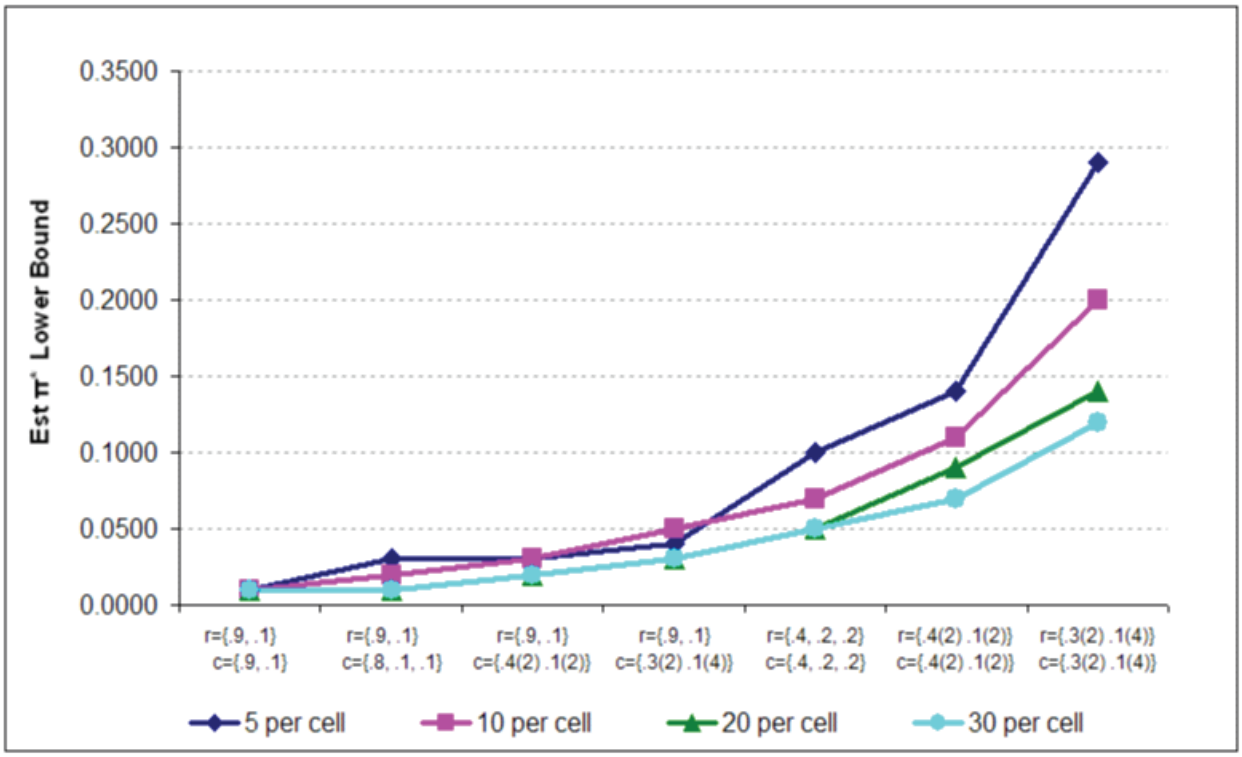


Figure 7: Empirical Simulation Based $\hat{\pi}_L$ with Extremely Distributed Marginals



As shown in Figures 8 and 9, for each size of contingency table, the lower bound estimate of $\hat{\pi}^*$ is generally smallest for extremely dispersed row and column marginal distributions, followed by slightly dispersed row and column marginal distributions; while largest for evenly distributed row and column tables. Different techniques for dealing with sampling zeros seem to have no effect on the lower bound estimate of $\hat{\pi}^*$ for either the RCL method or the true lower bound estimate based on empirical simulations.

Figure 8: $\hat{\pi}_L$ Comparison between the RCL Method and Empirical Simulation Method

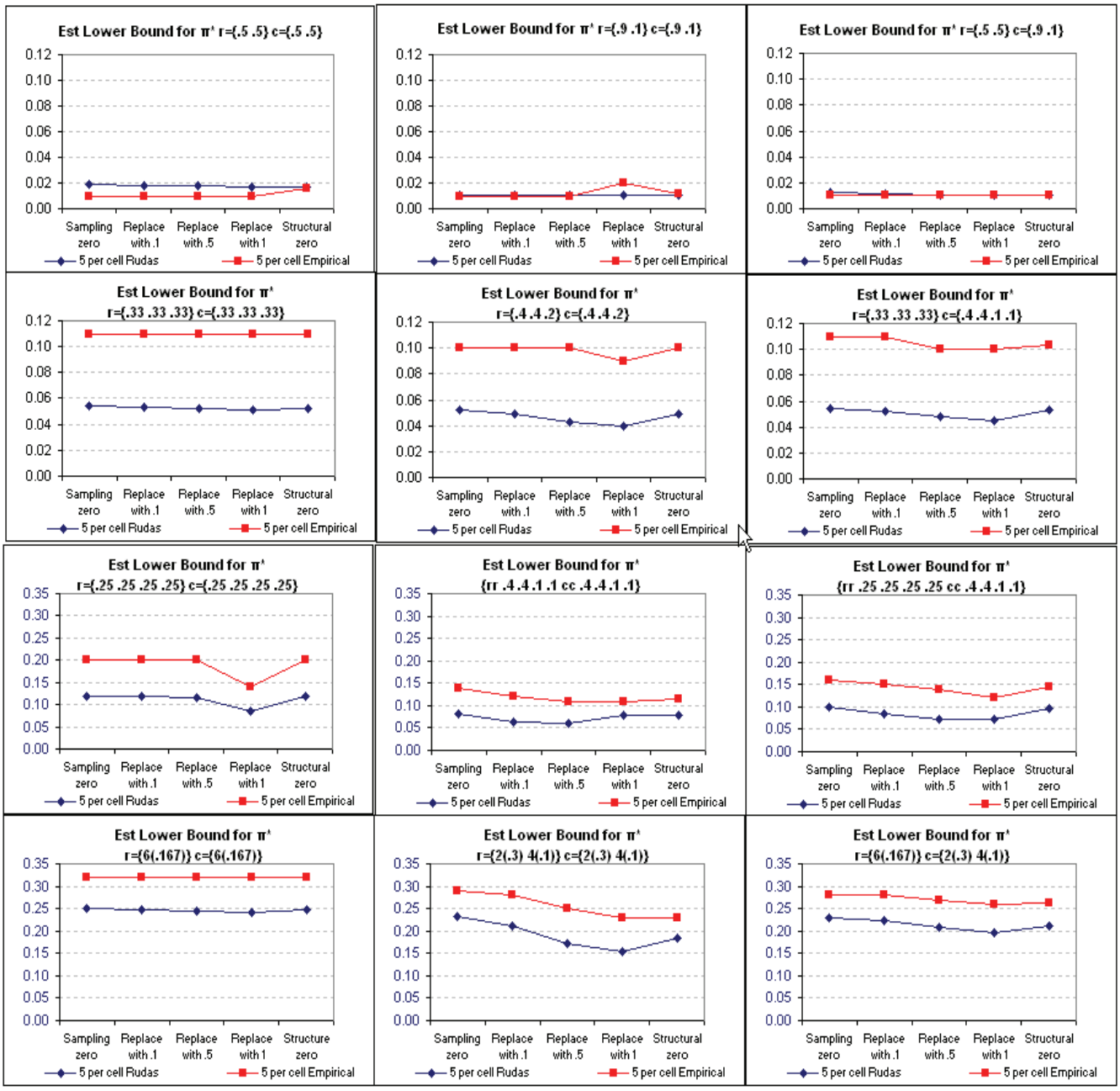
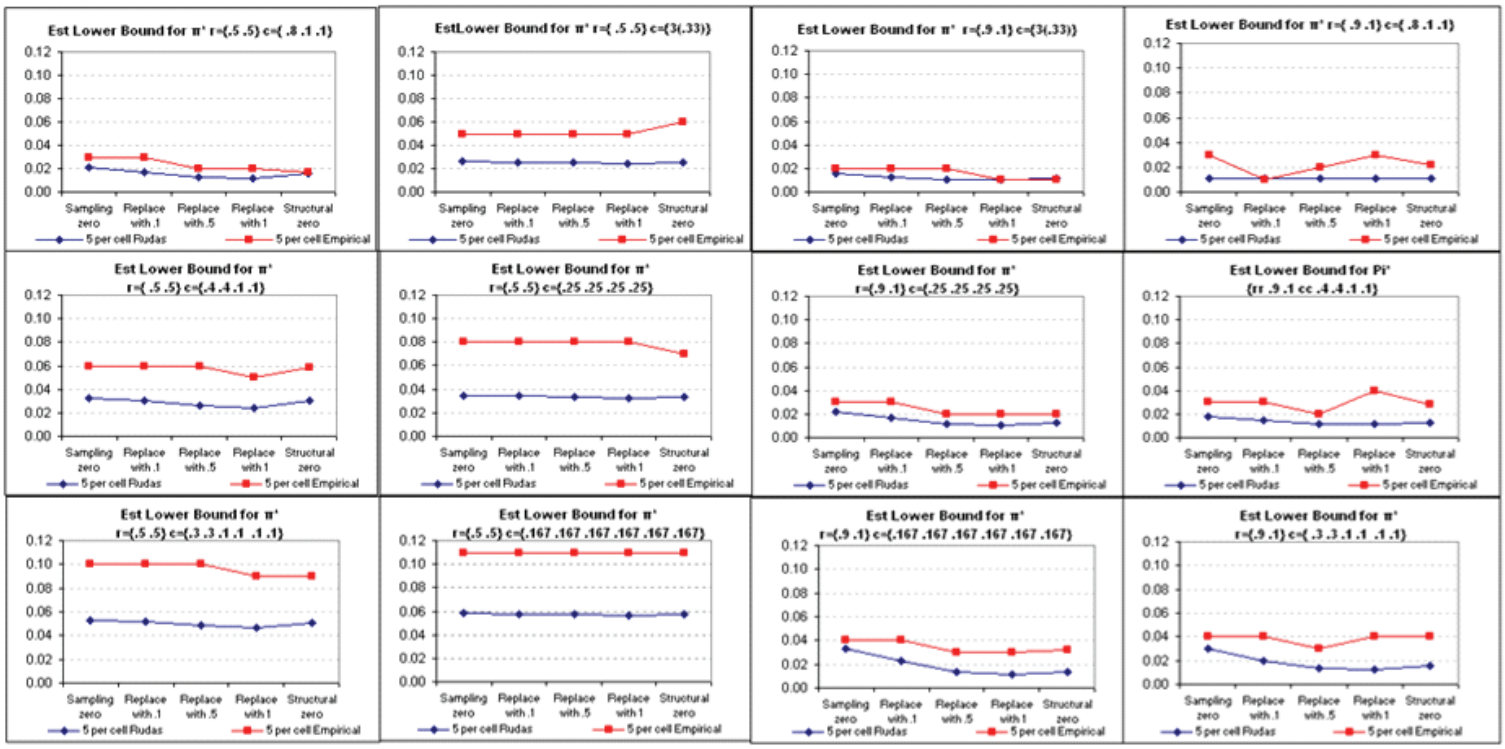


Figure 9: $\hat{\pi}_L$ Comparison between the RCL Method and Empirical Simulation Method (continued)



As shown in Figures 8 and 9, for each size of contingency table, the lower bound estimate of $\hat{\pi}^*$ is generally smallest for extremely dispersed row and column marginal distributions, followed by slightly dispersed row and column marginal distributions; while largest for evenly distributed row and column tables. Different techniques for dealing with sampling zeros seem to have no effect on the lower bound estimate of $\hat{\pi}^*$ for either the RCL method or the true lower bound estimate based on empirical simulations.

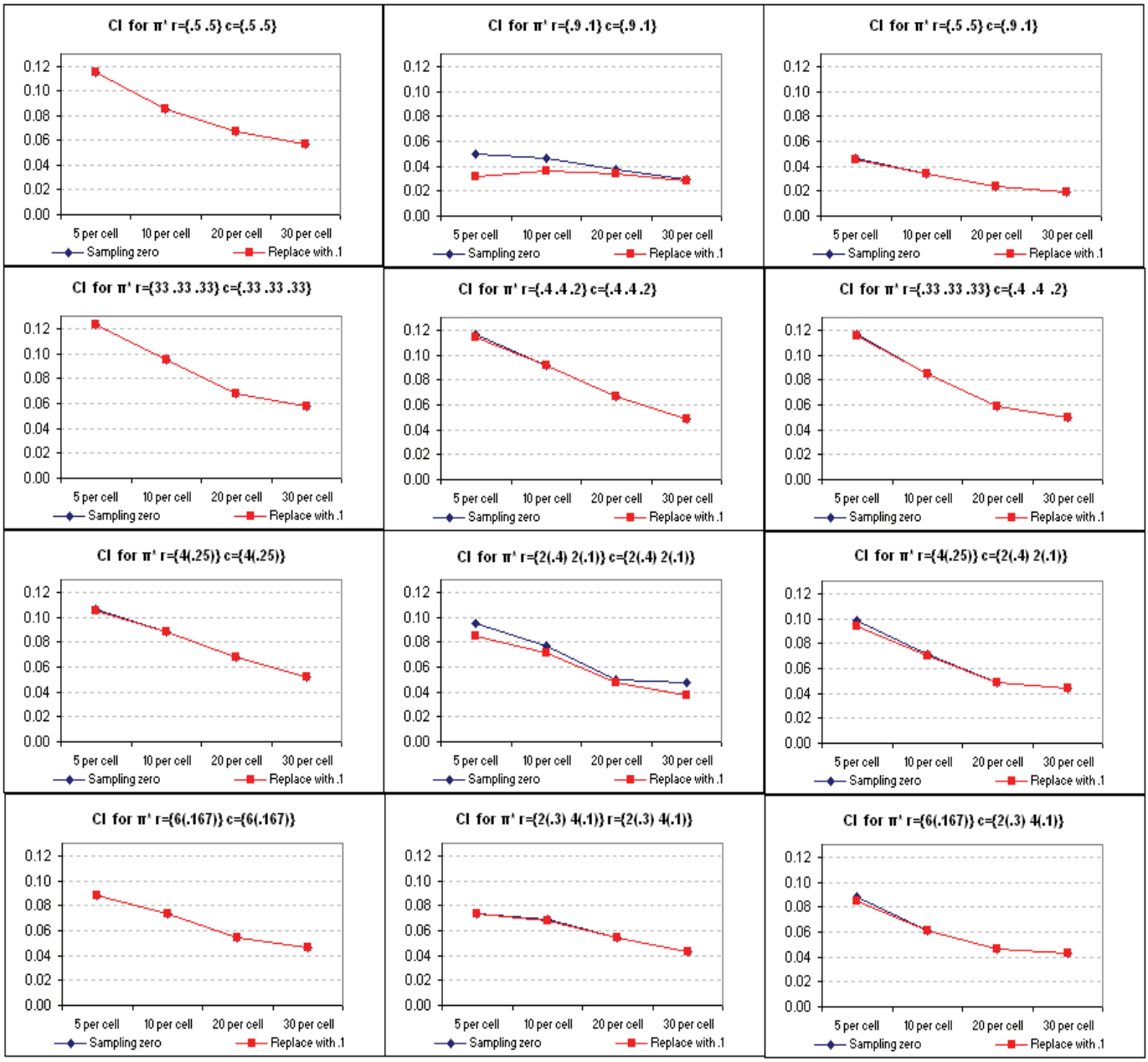
Confidence Interval and Standard Errors

Figures 10 and 11 show that, given the same table size, extremely dispersed row and column marginal distributions consistently provide narrower confidence intervals $\left(\hat{\pi}^* - \hat{\pi}_L\right)$ than evenly dispersed row and column tables using both the RCL method and empirical true estimates. Also, when sample size increases, confidence intervals become narrower for each table size and shrink approximately to the same confidence intervals for different marginal distribution for the same table size using both estimation methods. It is apparent that the RCL method underestimates the lower bound of $\hat{\pi}^*$ in many cases and, thus, leads to a higher standard error compared with empirical true lower bound estimates.

Example 1: Fatal Crashes by Speed Limit

Table 1 presents fatal crashes by speed limit and land use in the United States in 2004 from Traffic Safety Facts 2004, a compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. There are three categories in the land use variable (rural, urban and unknown), and six categories in the speed limit variable (30 mph or less, 35 or 40 mph, 45 or 50 mph, 55mph, 60 mph or higher and no statutory limit). This data table was used to compare the conclusion using traditional Chi-square and

Figure 10: Confidence Interval of $\hat{\pi}^*$ Following Empirical Simulation Method



related model fit methods and the mixture index of fit introduced by RCL. More specifically, compare different sampling zero techniques impact on $\hat{\pi}^*$ because there is one zero cell in the contingency table.

The value of the Pearson Chi-Square statistic is 7200.090, and the likelihood ratio, G^2 statistic is 7600.54 both with degrees of freedom equal to 10 ($P < 0.01$). Thus, an independence model is not tenable based on these Chi-squared tests of fit. As displayed in Table 3, the mixture index of fit $\hat{\pi}^*$ is 0.294, indicating that about 29.4% of the total of 37,295 cases (or, 10,965 cases) must be removed in order to attain perfect model fit. The mixture index of fit provides an interpretation consistent with traditional Chi-

Figure 11: Confidence Interval of $\hat{\pi}^*$ (continued)

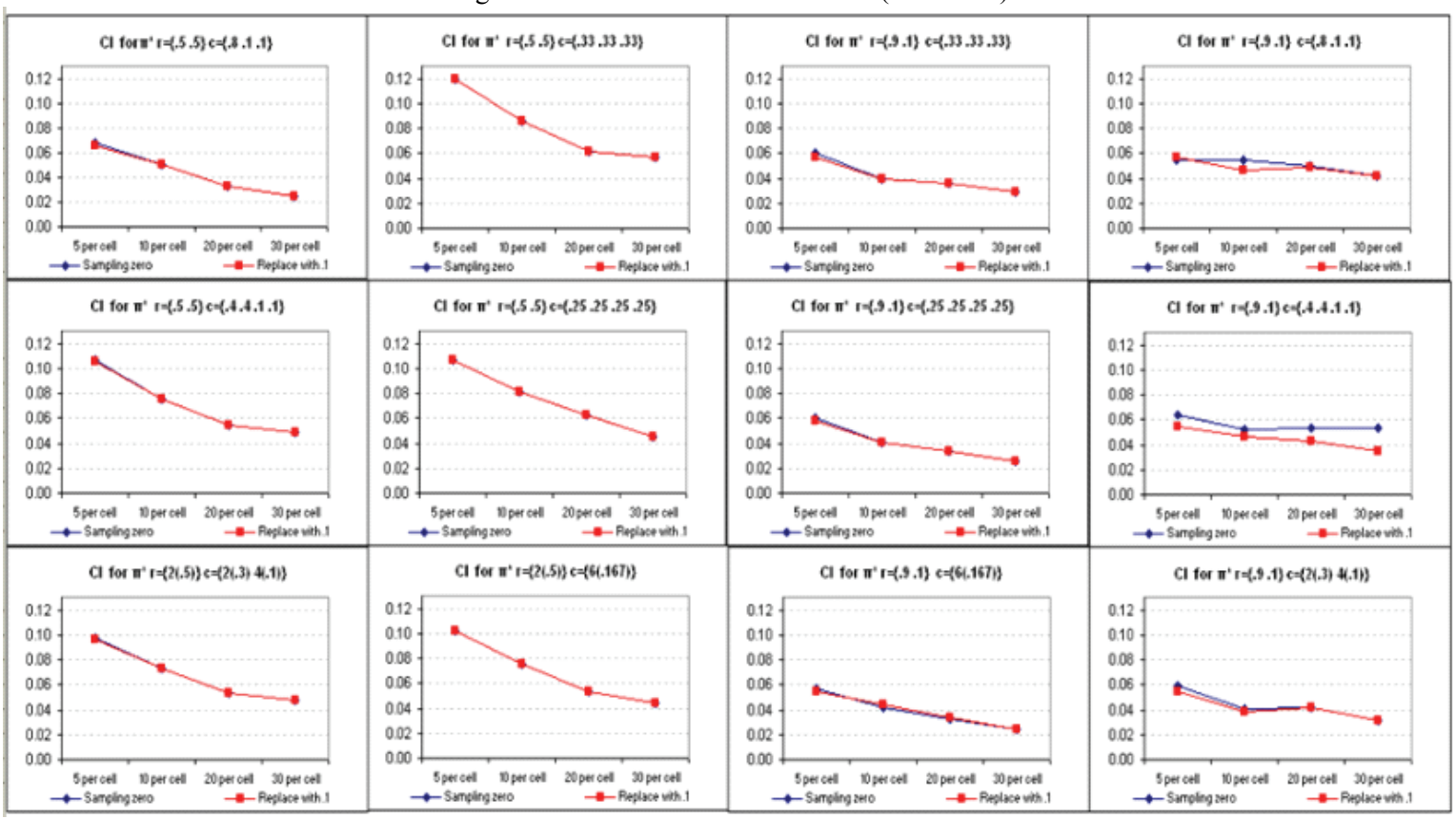


Table 1: Fatal Crashes by Speed Limit and Land Use

Speed Limit	Land Use		
	Rural	Urban	Unknown
30 mph or less	944	2929	27
35 or 40 mph	1951	4463	41
45 or 50 mph	3496	3559	46
55 mph	9646	2121	91
60 mph or higher	5484	2347	27
No statutory limit	92	31	0

Source: USDOT Traffic Safety Facts 2004 (Fatality Analysis Reporting System). Note: Omit 958 cases for the Unknown Speed Limit category.

Square analyses. Furthermore, $\hat{\pi}^*$ only decreases to 0.293 when replacing sample zero with the flattening constant 0.1 and further reduces to 0.291 when replacing with 0.5 and 1 as well as using the structural zero method. The amount of change in $\hat{\pi}^*$, as well as its 95% lower bound using different sampling zero techniques, is extremely small in this example. This occurs due to the very small percentage (0.62%) of unknown land. In fact, it would not substantially effect $\hat{\pi}^*$ even if the entire column were zeros.

Example 2: Eye Color and Hair Color

Table 2 presents a cross-classification of eye color and hair color table (Snee, 1974), a 4 × 4 table with total sample size of 592.

RCL utilized these data to study the properties of the mixture index of fit. In this study, these data were used to compare the differences in estimates that result from using sampling zeros and structural zeros. The 16 cells were set to zero one-by-one and the results are shown in Table 4. The percentage differences between use of sampling zero and structural zero techniques range from 11.1% to 40.0%, Note that 6 of these differences are statistically significant ($p < 0.05$) using conventional z tests for proportions. The largest reductions in $\hat{\pi}^*$ using structural zero occurs when black hair and hazel eye color is set to zero.

Table 2: Cross-classification of Eye Color and Hair Color

Eye Color	Hair Color			
	Black	Brunette	Red	Blonde
Brown	68	119	26	7
Blue	20	84	17	94
Hazel	15	54	14	10
Green	5	29	14	16

Source: Snee (1974), Diaconis & Efron (1985).

Table 3: Fit Statistics for Fatal Crashes by Speed Limit and Land Use

π	Chi-Square	G -Square	π Repl with .1	Chi-Square	G-Square	π Repl with .5	Chi-Square	G-Square	π Repl with 1	Chi-Square	G-Square
0	7200.090	7600.540	0.000	7287.906	7599.937	0.000	7287.426	7599.116	0.000	7287.426	7599.116
0.1	2947.100	3044.340	0.100	2946.904	3043.700	0.100	2946.353	3042.763	0.100	2946.353	3042.763
0.15	1527.210	1578.220	0.150	1527.010	1577.582	0.150	1526.444	1576.641	0.150	1526.444	1576.641
0.2	632.920	655.610	0.200	632.726	654.986	0.200	632.173	654.094	0.200	632.174	654.094
0.25	140.530	142.830	0.250	140.330	142.266	0.250	139.875	141.612	0.250	139.875	141.612
0.26	81.050	82.280	0.260	80.856	81.737	0.260	80.449	81.149	0.260	80.449	81.149
0.27	37.190	37.890	0.270	37.003	37.364	0.270	36.643	36.841	0.270	36.643	36.841
0.28	10.010	10.510	0.280	9.824	9.992	0.280	9.060	9.534	0.280	9.510	9.534
0.286	*2.100*	*2.560*	*0.286*	*1.923*	*2.058*	*0.285*	*2.498*	*2.501*	*0.285*	*2.498*	*2.501*
0.29	0.510	0.730	0.290	0.272	0.321	0.290	0.004	0.004	0.290	0.004	0.004
0.294	0.000	0.000	0.293	0.000	0.000	0.291	0.000	0.000	0.291	0.000	0.000
$\pi \geq 0.294$	***0.000***	***0.000***	***$\pi \geq 0.293$***	***0.000***	***0.000***	***$\pi \geq 0.291$***	***0.000***	***0.000***	***$\pi \geq 0.291$***	***0.000***	***0.000***

NOTE: 1. Italic figures denote a 95% lower bound for $\hat{\pi}^*$ and associated Chi-Square and G-Square values.

NOTE: 2. Italic and bolded figures denote $\hat{\pi}^*$ and the associated Chi-Square and G-Square values.

Table 4: $\hat{\pi}^*$ Comparison of Sampling Zero and Structure Zero using Eye Color Data
(Each cell manipulated to be zero in turn.)

Eye color	Hair color: Black			Brunette			Red			Blonde		
	Sampling zero	Structure zero	%	Sampling zero	Structure zero	%	Sampling zero	Structure zero	%	Sampling zero	Structure zero	%
Brown	0.32	0.27	18.52%	0.42	0.40	5.00%	0.37	0.32	15.63%	0.32	0.28	14.29%
Blue	0.37	0.31	**19.35%**	0.44	0.38	15.79%	0.38	0.31	**22.58%**	0.20	0.18	11.11%
Hazel	0.42	0.30	**40.00%**	0.38	0.32	**18.75%**	0.38	0.31	**22.58%**	0.32	0.30	6.67%
Green	0.35	0.29	**20.69%**	0.32	0.30	6.67%	0.34	0.30	13.33%	0.34	0.29	17.24%

NOTE: 1. Bold figures denote significant percentage difference ($p < .05$, conventional z test for proportions) between sampling zero and structural zero techniques.

Recommendations

Among all the sampling zero techniques compared in terms of parameter bias, replacing zeros with larger flattening constants such as 1 and the structural zero technique appear to perform better in the sense that, on average, $\hat{\pi}^*$ is smaller. Between these two techniques, the structural zero technique is generally recommended for extremely and slightly dispersed row and column marginal distributions tables with small sample sizes and small table sizes while in other cases replacing with larger flattening constant (i.e., 1) is preferred.

Based on the current findings, RCL standard error estimates were comparatively conservative. In general, it is preferable in practice to use variance estimates that tend to be conservative (i.e., larger) rather than liberal (i.e., smaller). However, it would be valuable to investigate the standard error of $\hat{\pi}^*$ using re-sampling methods to provide better guidance for users.

Implications for Future Research

1. Evenly distributed, slightly and extremely dispersed marginal distributions for each different size of tables were manipulated in the current study. It would be valuable to investigate more diversified marginal distribution in future studies.
2. As noted, the limitation of structural zero technique with number of zero cells might affect the results when compared with other sampling zero techniques. It would be of interest to investigate structural zero technique applied in two-point mixture model index in contingency tables with more than one zero when the independence assumption holds.
3. In order to attain reasonable execution times for the simulation, in this study, an increment of .01 was adopted to successively increase $\hat{\pi}^*$ when estimating $\hat{\pi}^*$ using an EM algorithm. For very small true values of π^*, it would be necessary to use a value of .001 or even .0001 in order to obtain a more detailed picture, especially for the lower bound of $\hat{\pi}^*$.
4. In a future study, it would be beneficial to investigate the standard error of $\hat{\pi}^*$ using other re-sampling methods (e.g., jackknife) and compare with RCL to provide a more concrete guide.
5. The larger value of flattening constants (e.g., 1) might affect the original data structural when sample size of a contingency table is small (e.g., 5 per cell) and thus the results could be slightly influenced. Alternative ways to define the flattening constants such as a percentage to total sample size is of interest in future study.
6. Finally, it would be valuable to evaluate the performance of π^* under conditions where the independence assumption does not hold.

References

Clogg, C. C., Rudas, T., & Xi, L. (1995). A new index of structure for the analysis of models for mobility tables and other cross-classifications. In P. Marsden (Ed.) *Sociological Methodology*, 197-222. Blackwell, Oxford.

Clogg, C. C., Rudas, T., & Matthews, S. (1997). Analysis of model misfit, structure, and local structure in contingency tables using graphical displays based on the mixture index of fit. In J. Blasius, M. Greenacre (Eds.) *Visualization of Categorical Data*, 425-439. New York: Academic Press.

Dayton, C. M. (1999). *Latent class scaling analysis*. Thousand Oaks, CA: Sage.

Dayton, C. M. (2003). Applications and computational strategies for the two-point mixture index of fit. *British Journal of Mathematical and Statistical Psychology*, *56*, 1-13.

Diaconis, P., & Efron, B. (1985). Testing for independence in a two-way contingency table: New interpretations of the Chi-square statistic. *The Annals of Statistics*, *13*, 845-874.

Fienberg, S. (1980). *The analysis of cross-classified categorical data* (*2nd Ed.*). Boston, MA: MIT Press.

Knoke, D., & Burke, P. J. (1980). *Log-linear models*. Newberry Park, California: Sage Publications, Inc.

National Highway Traffic Safety Administration. (2006). *Traffic safety facts, 2004*. Report No. DOT HS-809-919. Washington, DC: US Department of Transportation.

Rudas, T. (1998). The mixture index of fit. In Ferligoj (Ed.) *Advances in methodology, data analysis and statistics*, 15-22. University of Ljubljana.

Rudas, T. (1999). The mixture index of fit and minimax regression. *Metrika*, 50, 163-172.

Rudas, T., Clogg, C. C., Lindsay, & B. G. (1994). A new index of fit based on mixture methods for the analysis of contingency tables. *Journal of the Royal Statistical Society*, *B*, *56*(*4*), 623-639.

Rudas, T., & Zwick, R. (1997). Estimating the importance of differential item functioning. *Journal of Educational and Behavioral Statistics*, *22*, 31-45.

SAS Institute Inc. (2005). *SAS User's Guide, Version 9*. Cary, NC: SAS Institute Inc.

Snee, R. (1974). Graphical display of two-way contingency tables. *The American Statistician. 28*, 9-12.

Verdes, E., & Rudas, T. (2002). The π^* index as a new alternative for assessing goodness of fit of logistic regression. In: Haitovsky, y., Lerche, H. R., Ritov, y. (Eds.) *Foundations of statistical inference*, 167-177. New York: Springer.

Xi, L. (1994). *The mixture index of fit for the independence model in contingency tables*. Master of Arts paper, Department of Statistics, Pennsylvania State University

Xi, L., & Lindsay, B. G. (1996). A note on calculating the π^* index of fit for the analysis of contingency tables. *Sociological Methods & Research*, *25*, 248-259.

Zwick, R., Rhayer, D. T., & Wingersky, M. (1994). A simulation study of the methods for assessing differential item functioning in computerized adaptive tests. *Applied Psychological Measurement*, *189*, 121-140.